2022
THE JOURNAL ON TECHNOLOGY AND PERSONS WITH DISABILITIES

Caption UI/UX - Display Emotive and Paralinguistic Information in Captions

Joseph Mendis, Ramzy Oncy-Avila, Christian Vogler and Raja Kushalnagar
Gallaudet University
joseph.mendis@gallaudet.edu , ramzy.avila@gallaudet.edu , Christian.vogler@gallaudet.edu, raja.kushalnagar@gallaudet.edu

Abstract

Emotive and paralinguistic information are typically missing from conventional closed captions. To include some of this information in closed captions, we generated graphical representations of this information that is normally represented with descriptive words. We evaluate them in captions and report on how participants perceive them. These viewers then provided verbal and written feedback regarding positive and negative aspects of the various captions. We found that hard of hearing viewers were significantly more positive about this style of captioning than deaf viewers and that some viewers believed that these augmentations were useful and enhanced their viewing experience.

Journal on Technology and Persons with Disabilities
Santiago, J. (Eds): CSUN Assistive Technology Conference

Introduction

Captions are the overlaid text representation of speech or sound of a video. Captions allow viewers to follow the dialogue and the action of a program simultaneously. It provides access to the spoken content for people who cannot access the audio either because they cannot hear it, or the audio is not available. Captions can also provide information about who is speaking or about sound effects that might be important to understanding a news story, political event, or the plot. The importance of closed captions cannot be understated considering the millions of people who rely on them for watching videos (Crabb et al., 2015). The demand for captions will continue to rise as video content is being produced on the internet. YouTube has seen an increase in the number of hours of videos uploaded every minute from 500 as of May 2019 from 400 hours every minute in 2015 (Tankovska, 2021). Over the years, there have been many improvements due to regulatory requirements on providing guidelines for captioning, such as the United States of America's Federal Communication Committee (FCC), or the United Kingdom's Office of Communications (Ofcom) guidelines and regulations for closed captions on television.

Closed captioning has allowed people who are deaf and hard of hearing to be included as audience members. Many of these viewers have to deal with non-speech information and identifying sound sources. These issues challenge viewers to watch video and read captions that convey sufficient aural information yet remain synchronized and readable. For example, some DHH viewers miss the point of visual gags or fail to identify who is talking in a group, as some of the audio information such as music, sound effects, and speech prosody are not usually included in captioning. Due to the lack of standardization on how to represent non-speech information, sound information is usually limited to speech representation. Emotions are

expressed through a combination of verbal communication such as speech and its prosodic modifiers and paralinguistic characteristics such as facial expressions, gestures, gaze, body positions, and movements (Wang and Cheong, 2006). Much of the semantics of television and film are conveyed through the interactions among humans on-screen, background sound, and music. The literature on primitive emotions shows that there are at least six common emotions: anger, sadness, happiness, fear, surprise, and disgust (Ortony and Turner, 1990). A previously conducted study used avatars for speaker indicators in captions, which included an image of a character and their name beneath it, indicating the speaker and their position in the video (Fels & Vy, 2009).

Another important piece of information that is typically missing from closed captions are emotions. Closed captions are text-based, so the "expressions of paralinguistic and emotive information are typically missing in video media," (Rashid et al., 2006). Due to this, people who are deaf or hard-of-hearing are experiencing limited access to such video media.

It takes time to read and process captions, which are also subject to space and speed limitations. Viewers were generally satisfied with captioning quality for television, but were dissatisfied with missing words, spelling errors, and captions that moved too quickly caused dissatisfaction (Jensema, 1996, 1998). These studies also confirmed that reading speed and vocabulary levels limit the quantity and speed of text. There is often barely enough time and space within a caption to provide a verbatim translation of what is being said. Because of this, other non-verbal aspects central to the intended entertainment experience are omitted.

Viewers cannot simultaneously watch media and read the visual captions; instead, they have to switch between the two and inevitably lose information and context in watching the movies. In contrast, hearing viewers can simultaneously listen to the audio and watch the scenes.

Using compact representations such as graphics can help address the limited time spent on following the captions.

One of the difficulties of using graphics to convey information is deciding what information to convey and how best to convey it using the most appropriate graphics. Some solutions for this can be drawn from related fields, such as text communication. In messaging apps, such as iMessage, offer emojis, which are icons, mostly of faces that show various expressions for various emotions. Some systems allow users to add personal features to the emojis by merging a three-dimensional snapshot of their face with the emoji, which provide a personal touch to using the icon to express their emotions. Studies show that emojis have a positive effect on the ability of senders to communicate their message and of receivers to understand the message [Rovers and van Essen 2004]. Similar to using text to describe emotions in captions, using emojis to convey emotions requires a standardized lexicon.

Since there are many kinds of emotions and non-speech information, it is difficult to represent the range of possibilities with a limited set of symbols. A single non-speech sound can be captioned in multiple ways and there is no clear agreement on which one to pick. For example, a phone ringing could be represented in at least three different ways: [Phone Ringing] or as [Phone Rings Multiple Times] or [Phone rings 3 times].

Similarly, emojis and spatial information can be flexible enough to represent spatial information such as the location of the sound in a room. Attempting to express spatial information with text (as speech or in written form) is less efficient, more error-prone, and requires more descriptors and interpretations (Fels, 2001).

In this paper, we explore two simple but effective approaches to aid DHH viewers in watching movies with nonspeech audio content: emotive and paralinguistic information. Our

experiments measure students' preference and recall of captions and associated scenes in captioned media clips, by asking them to complete a survey after using each of our two tools compared to the baseline case of regular captions.

We discuss the relevant design criteria based on the feedback we received from users during our iterative design process and suggest future work that builds on these insights. The addition of compact visual information into caption can increase satisfaction and understanding of the captioned media related to traditional captions.

Methodology

The sample included 20 adult participants who are deaf or hard-of-hearing. A demographic questionnaire was designed for the study to gather information on our participants. This includes basic questions such as gender, age, ethnicity, education level, hearing level, sign language skill level, deaf identity, and their experiences using technology. Our participants all had prior experience with watching captioned movies from birth, reflecting the fact that they had grown up after the passage of the Americans with Disabilities Act. Participants that were included in our study were compensated $15 for their time in our 30-minute study sessions. The recruiting period occurred during the months of June and July 2021 in the United States.

We tested two conditions. Condition 1 is testing the use of Apple's feature called Memoji to display emotive information in captions. Condition 2 is testing the use of text description to display paralinguistic information. In condition 1 we tested the emotive information such as sad, happy, angry, confused, and fear. For condition 2, we tested paralinguistic information such as inspiration, sarcastic, quiet, surprised, and loud.

Table 1. Lists of All Types Under Each Condition: sad, happy, angry, confused, and fear for Memoji; and inspirational, sarcastic, quiet, surprised, and loud for paralinguistic information.

Condition 1: Memoji for Emotive Information	Condition 2: Text Description for Paralinguistic Information
Sad	Inspirational
Happy	Sarcastic
Angry	Quiet
Confused	Surprised
Fear	Loudly (screaming)

Results

Overall, there were some differences shown between the closed captions shown with emotive and paralinguistic information and the base closed captions that did not include either emotive or paralinguistic information.

After conducting the surveys, we asked participants their thoughts on such conditions to show either emotive or paralinguistic information, if they see a benefit from using such conditions, which condition they would prefer to use, and any feedback on what improvements they would like to see. We have gathered many different responses based on the follow up questions. In the table below, it shows that hard of hearing participants see a bigger improvement in both conditions than deaf participants as hearing capability might influence these conditions.

Table 2. Compares Responses Between Base and Emotion/Paralinguistic Enhanced Captions.

Difficulty Perceiving Emotion	Base	Condition	Difference
Deaf	2.96	2.4	0.56
HOH	2.92	2.2	0.72

Difficulty Understanding Emotion	Base	Condition	Difference
Deaf	3.24	2.56	0.68

Difficulty Understanding Emotion	Base	Condition	Difference
HOH	3.26	2.2	1.06

Difficulty Perceiving Paralinguistic Information	Base	Condition	Difference
Deaf	2.74	1.86	0.88
HOH	2.62	2.18	0.44

Difficulty Understanding Paralinguistic Information	Base	Condition	Difference
Deaf		1.84	1.06
HOH	3.06	1.88	1.18

Discussion

Our results indicated that providing participants with additional information within their captioning experience can be overall beneficial. Throughout most of our participant responses we saw that there was a consistent difference between the overall difficulty our participants experienced when watching the base captioning in comparison to our condition captioning styles: Memojis and Paralinguistic Information.

Many participants found the paralinguistic condition to be helpful and gave us common feedback to improve it such as moving it to the beginning of the caption instead of the ending so participants can understand the tone before reading the captions. Other feedback was given such as giving it more variations instead of having the information in all capitals by making it italicized or mixed capitals and have it more often for other speakers.

As for the Memoji condition, many participants liked it as a speaker identification and common feedback was given about it not having a lot of variations in facial expressions and that it can be visually distracting and hard to distinguish at a far distance. While the Memoji condition is a new concept to all the participants, it would probably take some practice in

learning how to receive information and get used to this new design.

Overall, most of the participants prefer the paralinguistic condition the most as it is the easiest for them to read and understand as some of them have seen similar designs on shows or movies that they watched. Overall, there were signs of improvement for both conditions in terms of understanding emotive and paralinguistic information in captions.

Conclusion

Our study shows that deaf and hard of hearing participants show improvement to understanding emotional and paralinguistic information in captions based on our added conditions to the captions. The ability to directly perceive salient environmental information adds a new dimension to accessibility in watching media and provides a deeper understanding. . Our findings indicate that these conditions are a good possible way to include such information, while we have received numerous responses and feedback for the conditions based on the follow-up questions response. Iterations of these conditions are needed to better improve the effectiveness of including emotive & paralinguistic information in captions. While combining both conditions in the same captions can be possible but it can be too much information to process, so considering a caption design that will reduce the cognitive load and make it easier to process the information would be the next step of research. We hope with this research, we help push out other ideas of including emotive or paralinguistic information in captions and help it become an option for users to use in mainstream captioning services.

Works Cited

Boyd, J., & Vader, E. (1972). Captioned Television for the Deaf. *American Annals of the Deaf, 117*(1), 34-37. Retrieved June 1, 2021, from http://www.jstor.org/stable/44389121

Burnham, D., Leigh, G., Noble, W., Jones, C., Grebennikov, M. T. L., & Varley, A. (2008). Parameters in Television Captioning for Deaf and Hard-of-Hearing Adults: Effects of Caption Rate Versus Text Reduction on Comprehension. *Journal of Deaf Studies and Deaf Education*, *13*(3), 391–404. https://doi.org/10.1093/deafed/enn003

Crabb, M., Jones, R., Armstrong, M., & Hughes, C. J. (2015). Online News Videos: The UX of Subtitle Position. *Proceedings of the 17th International ACM SIGACCESS Conference on Computers & Accessibility - ASSETS '15*. https://doi.org/10.1145/2700648.2809866

Jelinek Lewis, M. S., & Jackson, D. W. (2001, January 1). *Television Literacy: Comprehension of Program Content Using Closed Captions for the Deaf*. OUP Academic. https://academic.oup.com/jdsde/article/6/1/43/372192.

McKee, M. M., Paasche-Orlow, M. K., Winters, P. C., Fiscella, K., Zazove, P., Sen, A., & Pearson, T. (2015). Assessing Health Literacy in Deaf American Sign Language Users. *Journal of Health Communication*, *20*(sup2), 92–100. https://doi.org/10.1080/10810730.2015.1066468

Ortony, A. and Turner, T. J. 1990. What's basic about basic emotions? Psychological Review 97, 3, 315-331.

Rashid, R., Aitken, J., & Fels, D. I. (2006). Expressing Emotions Using Animated Text Captions. *Lecture Notes in Computer Science*, 24–31. https://doi.org/10.1007/11788713_5

Rashid, R., Vy, Q., & Fels, D. I. (2007). Dancing with Words. *Using Animated Text for Captioning, International Journal of Human–Computer Interaction*, 269–270.

Tankovska, H. (2021, January 26). *YouTube: hours of video uploaded every minute 2019*. Statista. https://www.statista.com/statistics/259477/hours-of-video-uploaded-to-youtube-every-minute/.

Television Decoder Circuitry Act. National Association of the Deaf. (n.d.). https://www.nad.org/resources/civil-rights-laws/television-decoder-circuitry-act/.

Vy, Q. V. (2012). Enhanced Captioning: Speaker Identification Using Graphical and Text-based Identifiers. https://doi.org/10.32920/ryerson.14652255.v1

Vy, Q. V., & Fels, D. I. (2009). Using Avatars for Improving Speaker Identification in Captioning. *Human-Computer Interaction – INTERACT 2009*, 916–919. https://doi.org/10.1007/978-3-642-03658-3_110

Wang, H. L, and Cheong, L.F. (2006). Affective understanding in film. IEEE Transactions on Circuits and Systems for Video Technology 16, 6, 689-704.